Causal Convolutional Encoder Decoder-Based Augmented Kalman Filter for Speech Enhancement

Sujan Kumar Roy, Kuldip K. Paliwal
Signal Processing Laboratory, Griffith School of Engineering
Griffith University, Brisbane, QLD, Australia, 4111
sujankumar.roy@griffithuni.edu.au, k.paliwal@griffith.edu.au

***Abstract*—Speech enhancement using augmented Kalman filter (AKF) suffers from the biased estimates of the linear prediction coefficients (LPCs) of speech and noise signal in noisy conditions. The existing AKF was particularly designed to enhance the colored noise corrupted speech. In this paper, a causal convolutional encoder-decoder (CCED)-based method utilizes the LPC estimates of the AKF for speech enhancement. Specifically, a CCED network is used to estimate the instantaneous noise spectrum for computing the LPCs of noise on a framewise basis. Each noise corrupted speech frame is pre-whitened by a whitening filter, which is constructed with the noise LPCs. The speech LPCs are computed from the pre-whitened speech. The improved speech and noise LPCs enables the AKF to minimize the residual noise as well as distortion in the enhanced speech. Objective and subjective testing on NOIZEUS corpus reveal that the enhanced speech produced by the proposed method exhibits higher quality and intelligibility than the benchmark methods in various noise conditions for a wide range of SNR levels.**

***Index Terms*—Speech enhancement, augmented Kalman filter, convolution neural network, LPC, whitening filter.**

I. Introduction

The objective of a speech enhancement algorithm (SEA) is to estimate the clean speech from the noisy speech signal. The SEAs can be used as a pre-processor for many speech processing systems, such as voice communication systems, hearing-aid devices, speech recognition. Various SEAs, such as spectral subtraction (SS) [1], [2], MMSE [3], [4], Wiener Filter (WF) [5], [6], Kalman filter (KF) [7] have been introduced over the decades. However, it is still a demanding work to develop an efficient SEA for real-world noise conditions.

The SS-based SEA heavily depends on the accuracy of noise power spectral density (PSD) estimates [8]. The under/over-estimation of the noise PSD introduces musical noise and distortion in the enhanced speech [9, Chapter 5]. The performance of the MMSE and WF based SEA somehow depends on the accuracy of the *a priori* SNR estimates in practice. In [3], Ephraim and Malah proposed a decision-directed (DD) approach to compute the *a priori* SNR in noisy conditions. However, this approach uses the speech and noise power spectrum estimated from the previous noisy speech frame, leading to an inaccurate estimate of the *a priori* SNR for the current frame. The biased estimate of the *a priori* SNR in the MMSE-based SEA typically introduce musical noise and spectral distortion in the enhanced speech [9].

The efficiency of KF-based SEA depends on how accurately the key parameters, LPCs are estimated in noisy conditions. Paliwal and Basu for the first time introduced KF-based SEA for enhancing stationary noise corrupted speech [7]. However, the LPCs are computed from the clean speech signal, which is unavailable in practice. In [10], Gibson *et al.* introduced an augmented KF (AKF) for enhancing colored noise corrupted speech. In this method, the LPC estimates for the current noisy speech frame are computed from the filtered signal of the previous iteration by AKF. Although the enhanced speech (after 2-3 iterations) shows SNR improvement, however, suffering from spectral distortion as well as musical noise. In [11], Roy *et al.* proposed a sub-band iterative KF-based SEA. Due to processing the high-frequency sub-bands (SBs) among the 16 decomposed SBs for a given noise corrupted utterance, some noise components may still remain in the low-frequency SBs. The enhanced speech also suffers from distortion. In [12], George *et al.* introduced a robustness metric-based tuning of the AKF. This SEA is particularly designed for colored noise suppression. In addition, the robustness metric-based tuning of the AKF gain causes distortion in the enhanced speech.

Over the decades, the deep neural network (DNN) has been used widely for speech enhancement [13]. The DNN usually gives an estimate of the ideal binary mask (IBM), which is used to compute the clean speech spectrum [13]. It is shown that the ideal ratio mask (IRM) [14] exhibits better speech quality than the IBM. In [15], Williamson et al. introduced a complex ideal ratio mask (cIRM), which is capable to recover both the amplitude and phase spectrum of clean speech. However, the masking technique usually introduces musical noise in the enhanced speech [14].

In [16], a convolutional encoder decoder (CED)-based SEA has been proposed. It was particularly designed to enhance the babble noise corrupted speech. In [17], a long short-term memory (LSTM) was incorporated with a CED to form a convolutional recurrent network (CRM) for speech enhancement. The CRM network [17] is constructed with 2D Convolution (Conv2D) layers, which is normally required for processing image data. Since speech signal is 1D, it can be processed with 1D convolution (Conv1D) layer as used in CED [16]. Thus, CRM [17] takes huge training parameters, which increases the training time accordingly. In [18], a fully convolutional neural network (FCNN)-based SEA has been introduced. It processes the raw-waveform of noise corrupted speech, yielding an estimate of clean speech waveform. Thus, the enhanced speech does not depend on the phase spectrum,

which has a significant impact on other acoustic-domain SEAs [13], [14], [16] (keep the phase-spectrum unprocessed). In [19], Zheng et al. introduced a phase-ware SEA using DNN. Here, the phase information (converted to the instantaneous frequency deviation (IFD)) is jointly used with different masks, namely ideal amplitude mask (IAM) as a training target. The clean speech spectrum is reconstructed with the estimated mask and the phase information (extracted from the IFD).

Yu *et al.* introduced a KF-based SEA, where the LPCs are estimated using a traditional DNN [20]. However, the training is performed with four different noise recordings including four SNR levels. Technically, it reduces the performance of this SEA for a wide range of noise conditions as well as SNR levels. Also, the noise covariance is estimated from the initial frames of noisy speech (considered as silent), which is irrespective with the non-stationary noise conditions.

The direct estimation of speech spectrum using benchmark deep learning methods reported in literature may suffer from musical noise and distortion. Our investigation reveals that the estimate of noise spectrum using deep learning technique would be more beneficial, since it is a crucial parameter for most of the SEAs in literature. For example, the AKF-based SEA suffering from the noise LPC estimates in practice. In this paper, a causal convolutional encoder-decoder (CCED) network addresses the speech and noise LPC estimates of the AKF for speech enhancement. Specifically, the CCED network gives an estimate of the instantaneous noise spectrum for computing the noise LPCs on a framewise basis. A whitening filter is then constructed with the noise LPCs to pre-whiten the noise corrupted speech frame prior to estimate speech LPCs. With the improvement of speech and noise LPCs, the AKF is found to be effective in minimizing the residual noise as well as distortion in the enhanced speech. The efficiency of the proposed SEA is compared against the benchmark SEAs using objective and subjective testing on NOIZEUS corpus.

II. AKF FOR COLORED NOISE SUPPRESSION

Assuming the colored noise $v(n)$ to be additive with speech $s(n)$ and uncorrelated each other, at sample n, the noisy speech $y(n)$ is given by:

$$y(n) = s(n) + v(n). \tag{1}$$

The $s(n)$ and $v(n)$ of eq. (1) can be modeled with p^{th} and q^{th} order linear predictors as [21]:

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + w(n), \tag{2}$$

$$v(n) = -\sum_{j=1}^{q} b_j v(n-j) + u(n), \tag{3}$$

where $\{a_i; i = 1, 2, \ldots, p\}$ and $\{b_j; j = 1, 2, \ldots, q\}$ are the LPCs, $w(n)$ and $u(n)$ are assumed to be white noise with zero mean and variance σ_w^2 and σ_u^2, respectively.

Eqs. (1)-(3) can be used to form the following augmented state-space model (ASSM) of AKF as [12]:

$$\boldsymbol{x}(n) = \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \boldsymbol{d}\boldsymbol{z}(n), \tag{4}$$

$$y(n) = \boldsymbol{c}^T\boldsymbol{x}(n). \tag{5}$$

In the above ASSM,

1) $\boldsymbol{x}(n) = [s(n) \ \ldots s(n-p+1)\, v(n) \ \ldots \ v(n-q+1)]^T$ is a $(p+q)\times 1$ state-vector,
2) $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_s & 0 \\ 0 & \boldsymbol{\Phi}_v \end{bmatrix}$ is a $(p+q)\times(p+q)$ state-transition matrix with:

$$\boldsymbol{\Phi}_s = \begin{bmatrix} -a_1 & -a_2 & \ldots & a_{p-1} & a_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

$$\boldsymbol{\Phi}_v = \begin{bmatrix} -b_1 & -b_2 & \ldots & b_{q-1} & b_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

3) $\boldsymbol{d} = \begin{bmatrix} \boldsymbol{d}_s & 0 \\ 0 & \boldsymbol{d}_v \end{bmatrix}$, where $\boldsymbol{d}_s = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^\top$, $\boldsymbol{d}_v = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^\top$,
4) $\boldsymbol{z}(n) = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}$,
5) $\boldsymbol{c}^T = \begin{bmatrix} \boldsymbol{c}_s^T & \boldsymbol{c}_v^T \end{bmatrix}$, where $\boldsymbol{c}_s = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T$ and $\boldsymbol{c}_v = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T$ are $p\times 1$ and $q\times 1$ vectors,
6) $y(n)$ is the noisy measurement at sample n.

Firstly, $y(n)$ is windowed into non-overlapped and short (e.g., 20 ms) frames. For a particular frame, the AKF computes an unbiased and linear MMSE estimate, $\hat{\boldsymbol{x}}(n|n)$ at sample n, given $y(n)$ by using the following recursive equations [12]:

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{\Phi}\hat{\boldsymbol{x}}(n-1|n-1), \tag{6}$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^T + \boldsymbol{d}\boldsymbol{Q}\boldsymbol{d}^T, \tag{7}$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}(\boldsymbol{c}^T\boldsymbol{\Psi}(n|n-1)\boldsymbol{c})^{-1}, \tag{8}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n-1)], \tag{9}$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\boldsymbol{\Psi}(n|n-1), \tag{10}$$

where $\boldsymbol{Q} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix}$ is the process noise covariance.

For a noisy speech frame, the error covariances $\boldsymbol{\Psi}(n|n-1)$ and $\boldsymbol{\Psi}(n|n)$ corresponding to $\hat{\boldsymbol{x}}(n|n-1)$ and $\hat{\boldsymbol{x}}(n|n)$, and the Kalman gain $\boldsymbol{K}(n)$ are continually updated on a samplewise basis, while ($\{a_i\}$, σ_w^2) and ($\{b_k\}$, σ_u^2) remain constant. At sample n, $\boldsymbol{g}^\top\hat{\boldsymbol{x}}(n|n)$ gives the estimated speech, $\hat{s}(n|n)$, where $\boldsymbol{g} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix}^\top$ is a $(p+q)\times 1$ column vector. As in [12], $\hat{s}(n|n)$ is given by:

$$\hat{s}(n|n) = [1 - K_0(n)]\hat{s}(n|n-1) + K_0(n)[y(n) - \hat{v}(n|n-1)], \tag{11}$$

where $K_0(n)$ is the 1^{st} component of $\boldsymbol{K}(n)$, given by [12]:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \beta^2(n) + \sigma_u^2}, \tag{12}$$

where $\alpha^2(n)$ and $\beta^2(n)$ are the transmission of *a posteriori* error variances (of the speech and measurement noise samples) by the augmented dynamic model from the previous time sample, $n-1$ [12].

Eq. (11) reveals that $K_0(n)$ has a significant impact on $\hat{s}(n|n)$ estimates (the output of the AKF). In practice, the poor estimates of $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ introduce bias in $K_0(n)$, which affects the estimates of $\hat{s}(n|n)$. In the proposed SEA, a CCED network utilizes the speech and noise LPC estimates of the AKF, leading to an improved $\hat{s}(n|n)$ estimate.

III. Proposed Speech Enhancement System

Fig. 1 shows the block diagram of the proposed SEA. Firstly, a 32 ms rectangular window with 50% overlap was considered for converting $y(n)$ (eq. (1)) into frames $y(n,l)$, i.e., $y(n,l) = s(n,l) + v(n,l)$, where $l\epsilon\{0,1,2,\ldots,N-1\}$ is the frame index with N being the total number of frames in an utterance, and M is the total number of samples within each frame, i.e., $n\epsilon\{0,1,2,\ldots,M-1\}$. The DFT coefficients

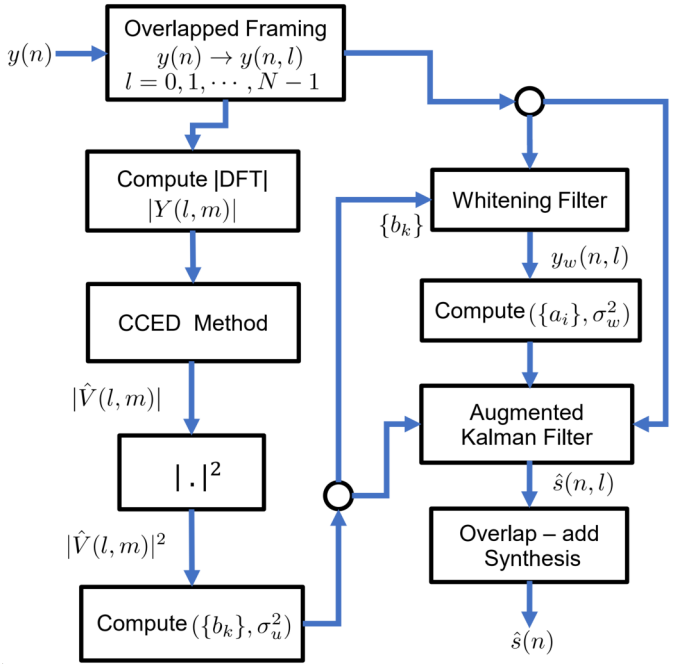

Fig. 1. Block diagram of the proposed SEA.

$Y(l,m)$, $S(l,m)$, and $V(l,m)$ are found using the Hamming window and correspond to $y(n)$, $s(n)$ and $v(n)$. These can also be represented as:

$$Y(l,m) = S(l,m) + V(l,m), \tag{13}$$

where m is the discrete-frequency index.

It is assumed that $S(l,m)$ and $V(l,m)$ follow a Gaussian distribution with zero-mean and variances; $E\{|S(l,m)|^2\} = \lambda_s(l,m)$, and $E\{|V(l,m)|^2\} = \lambda_v(l,m)$, where $E\{\cdot\}$ represents the statistical expectation operator.

A. Proposed $(\{b_k\}, \sigma_u^2)$ and $(\{a_i\}, \sigma_w^2)$ Estimation Method

The existing AKF-based SEA [12] estimates the noise from the initial noise corrupted speech frames by considering that there remains no speech. Then compute $(\{b_k\}, \sigma_u^2)$ from the estimated noise, which remains constant during processing all the frames for a given noise corrupted speech utterance. This concept may be effective for enhancing the colored noise corrupted speech. Due to the time varying nature of non-stationary noise amplitude, it is required to update $(\{b_k\}, \sigma_u^2)$ continuously during processing each noise corrupted speech frame. Thus, $(\{b_k\}, \sigma_u^2)$ estimation process in [12] becomes irrespective with the non-stationary noise conditions.

In the proposed SEA, we introduce a CCED-based method (described in section III-B) to estimate the instantaneous noise spectrum, $|\widehat{V}(l,m)|$ for a given noisy speech spectrum, $|Y(l,m)|$ on a framewise basis. By taking square of $|\widehat{V}(l,m)|$, i.e., $|\widehat{V}(l,m)|^2$, we get the instantaneous noise PSD from where $(\{b_k\}, \sigma_u^2)$ are computed. Specifically, the |IDFT| of $|\widehat{V}(l,m)|^2$ yields an estimate of the noise autocorrelation, $\widehat{R}_{vv}(\tau)$, where τ is the autocorrelation lag. By solving $\widehat{R}_{vv}(\tau)$ using the Levinson-Durbin recursion [21], the $(\{b_k\}, \sigma_u^2)$ $(q = 20)$ estimates are obtained. Then $\{b_k\}$'s are used to design the whitening filter, $H_w(z)$ as [21]:

$$H_w(z) = 1 + \sum_{k=1}^{q} b_k z^{-k}. \tag{14}$$

Employing $H_w(z)$ to $y(n,l)$ gives the pre-whitened speech, $y_w(n,l)$. Then $(\{a_i\}, \sigma_w^2)$ $(p = 10)$ are computed from $y_w(n,l)$ using autocorrelation method [21].

B. CCED for Noise Spectrum Estimation

We propose a CCED-based method to estimate $|\widehat{V}(l,m)|$. The proposed CCED network structure is shown in Fig. 2. It consists of a convolution encoder followed by a corresponding decoder. The encoder consists of a stack of five convolution layers. Unlike 2-dimensional convolution layers (Conv2D) in [17], we have used 1-dimensional convolutional layer (Conv1D), since it is appropriate to process the 1D speech signal. The Conv1D layer also reduces huge training parameter as well training time. The decoder also consists of a stack of five Conv1D layers. In addition, we have used the causal Conv1D layer [22]. Fig. 3 demonstrates the operating principle of the standard and causal Conv1D layers. The standard Conv1D layers (Fig. 3 (a)) are comprised of filters that capture the local correlation of nearby data points, thus leaking the future information into the current data during operating. Conversely, in the causal Conv1D layer (Fig. 3 (b)), the output at any time step t only uses the information from the previous time steps, i.e., 0 to $t-1$ [22]. It allows the CCED network for real-time noise spectrum estimation.

The CCED network maps the single-sided magnitude spectrum (257-point DFT coefficients including the Nyquist frequency components) of noisy speech, $|Y(l,m)|$ to that of the

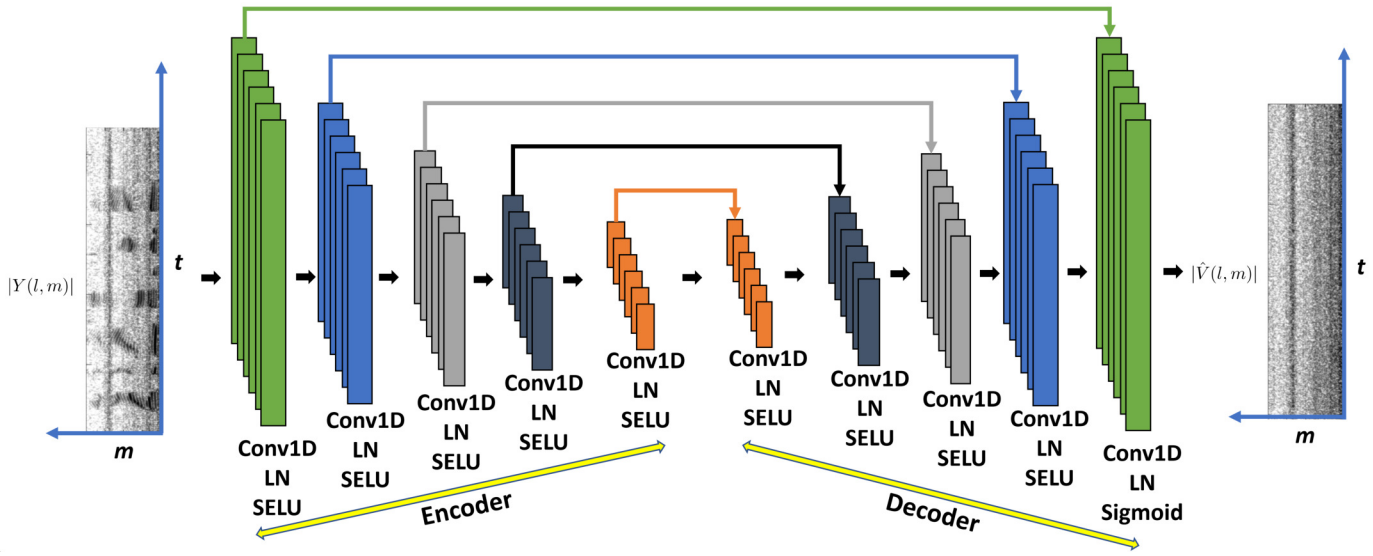

Fig. 2. Architecture of the proposed CCED network for noise spectrum estimation.

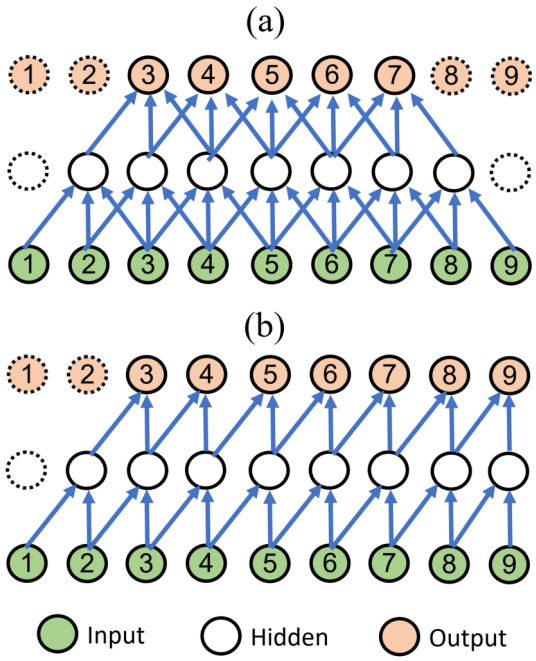

Fig. 3. Working principle of: (a) standard and (b) causal Conv1D layer.

noise spectrum, $|\widehat{V}(l,m)|$. Therefore, the output size of the first Conv1D layer in the encoder is 257. Specifically, the output size of the Conv1D layers in the encoder is decreased in the order of 257, 128, 64, 32, 16, which is increased in the decoder Conv1D layers, i.e., 16, 32, 64, 128, 257. We also use symmetric kernel in each Conv1D layer. The kernel size in the encoder Conv1D layers is increased gradually according to 1, 3, 5, 7, 9, which is decreased in the order of 9, 7, 5, 3, 1 for the decoder Conv1D layers. Therefore, the proposed CCED network encodes the features into lower dimension along the encoder and achieves decompression along the decoder. Each of the encoder-decoder layer is passed through a layer normalization (LN) [23] followed by SELU activation function [24], except the last layer, which passes through a sigmoid activation function [25] as it is the output layer. Reason of using SELU activation is that it has less impact on vanishing gradients than that of ReLU [26] and ELU [27]. Also, SELUs itself learn faster and better than ReLU and ELU even if they are combined with layer normalization [24]. Unlike [17], the Conv1D layer in the CCED network makes pooling and up-sampling unnecessary in the encoder and decoder layers.

To improve the flow of information and gradients throughout the network, we also utilize skip connection between the causal Conv1D usits of encoder and decoder. It resolves the so called vanishing gradient issue in a deep neural network. The skip connection is represented by arrows (used same color for the corresponding Conv1D units) as shown in Fig. 2.

IV. Speech Enhancement Experiment

A. Training Set

For training the proposed CCED network, a total of $30,000$ clean speech recordings are randomly selected belonging to the *train-clean-100* set of the Librispeech corpus [28] ($28,539$), the CSTR VCTK corpus [29] ($42,015$), and the si^* and sx^* training sets of the TIMIT corpus [30] ($3,696$). The 5% of $30,000$, i.e., 1500 speech recordings are randomly selected for cross-validation of the CCED network accuracy during training. Thus, $28,500$ speech recordings are used for training of the CCED network. Also, a total of 500 noise recordings are randomly selected from the QUT-NOISE dataset [31], the Nonspeech dataset [32], the Environmental Background Noise dataset [33], [34], the noise set from the MUSAN corpus [35]. The 5% of 500, i.e., 25 noise recordings are selected for cross-validation purposes, while the remaining 475 of them are used for training. All the clean speech and noise recordings are single-channel, with a sampling frequency of 16 kHz.

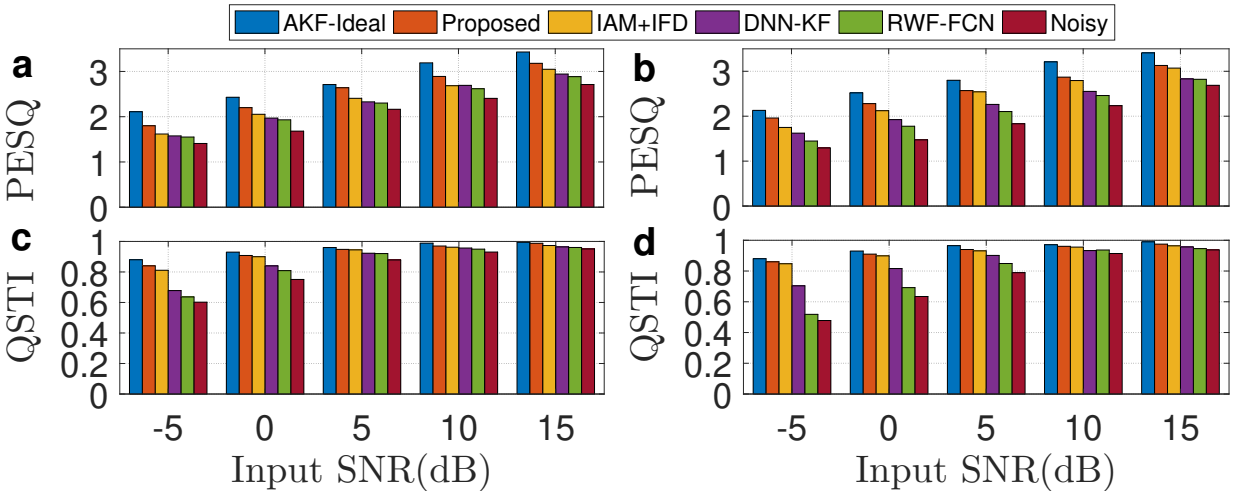

Fig. 4. Performance comparison of the proposed SEA with the benchmark SEAs in terms of the average: PESQ; (a) *passing car*, (b) *restaurant* and QSTI; (c) *passing car*, (d) *cafe babble* noise conditions.

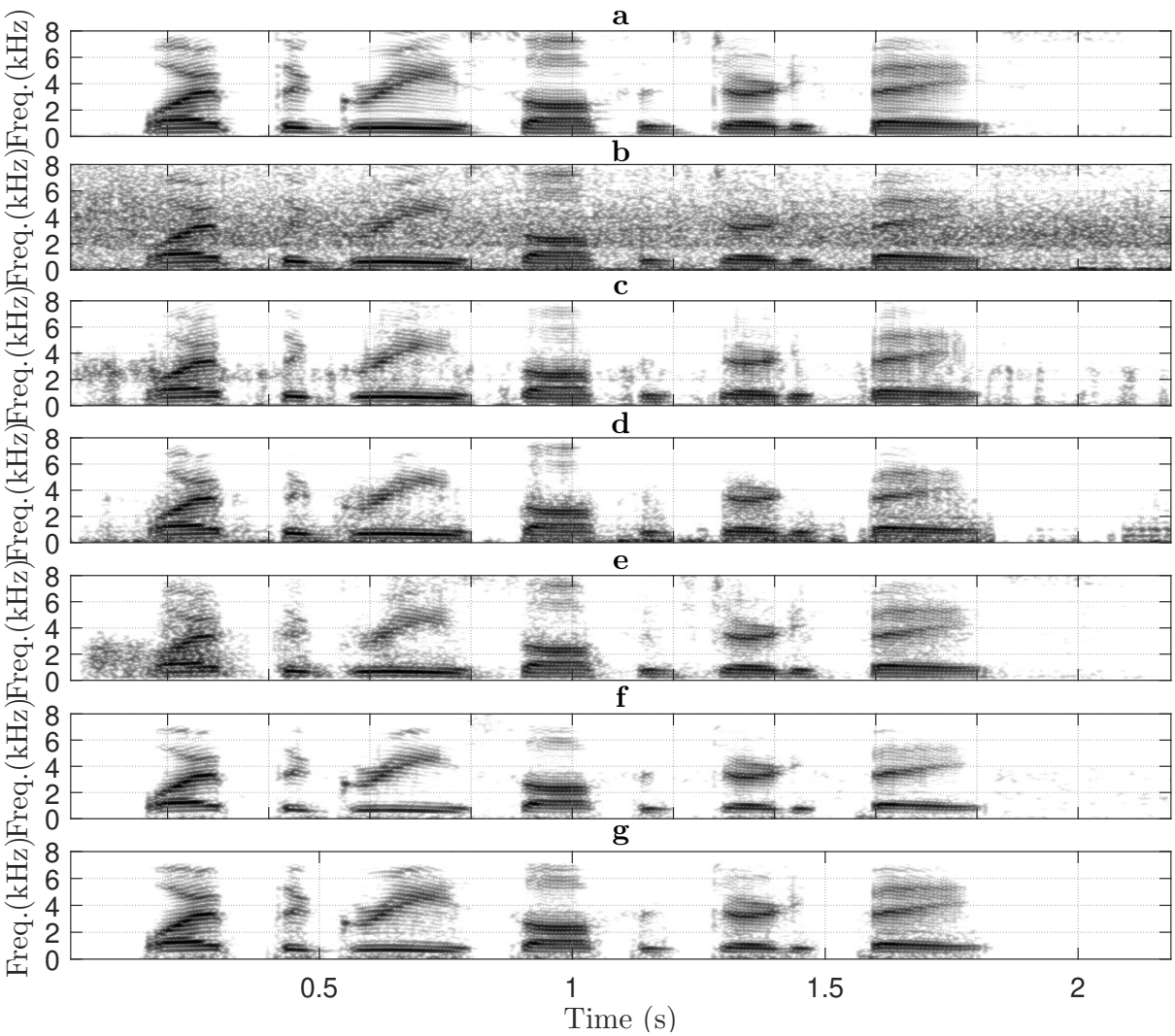

Fig. 5. (a) Clean speech, (b) noisy speech (sp05 is corrupted with 5 dB passing car noise), the enhanced speech spectrograms produced by the: (c) RWF-FCN [18], (d) DNN-KF [20], (e) IAM+IFD [19], (f) proposed, and (g) AKF-Ideal methods.

B. Training Strategy

The following training strategy was employed to train the proposed CCED network for noise spectrum estimation:

- The 'mean square error' is chosen as the loss function.
- The *Adam* algorithm [36] with default hyperparameters is also selected for gradient descent optimisation.
- Gradients are clipped between $[-1, 1]$.
- 120 epochs are used to train the CCED network.
- The number of training examples in an epoch is equal to the number of clean speech recordings used in the training set $(28,500)$.
- The noisy speech signals are generated as follows: each randomly selected clean speech recording (without replacement) is corrupted with a randomly selected noise recording (without replacement) at a randomly selected SNR level (-10 to +20 dB, in 1 dB increments).

C. Test Set

For objective experiments, 30 clean speech utterances belonging to six speakers (3 male and 3 female) are taken from the NOIZEUS corpus. The speech recordings are sampled at 16 kHz [9, Chapter 12]. We generate a noisy speech data set by corrupting the speech recordings with (*passing car*) and (*cafe babble*) noise recordings selected from [33], [34] at multiple SNR levels varying from -5dB to +15 dB, in 5 dB increments. It is important to note that both the speech and noise recordings are unseen and not used in training the CCED network.

D. Evaluation Metrics

The objective quality and intelligibility evaluation are carried out through the perceptual evaluation of speech quality (PESQ) [37] and quasi-stationary speech transmission index (QSTI) [38] measures. We also analyze the spectrograms of the enhanced speech produced by the proposed and benchmark SEAs to quantify the level of residual noise and distortion.

The subjective evaluation was carried out through blind AB listening test [39, Section 3.3.4]. It is conducted on the utterance sp05 ("*Wipe the grease off his dirty face*") corrupted with 5 dB *passing car* noise. The enhanced speech produced by five SEAs as well as the corresponding clean and noisy speech recordings, a total of 42 stimuli pairs played in a random order to each listener, excluding the comparisons between the same method. For each pair, the listener prefers the first or second stimuli which is perceptually better, or a third response indicating no difference is found between them. A 100% award is given to the preferred method, 0% to the other, and 50% to each method for the similar preference response. Participants could re-listen to stimuli if required. Five English speaking listeners participate in the AB listening tests. The average of the preference scores given by the listeners, termed as the mean preference score (%).

The performance of the proposed method is carried out by comparing it with the benchmark methods, such as raw waveform processing using FCNN (RWF-FCN) method [18], phase-aware DNN (IAM+IFD) method [19], deep learning-based KF (DNN-KF) method [20], AKF-Ideal method (where $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ are computed from the clean speech and noise signal) and Noisy (noise corrupted speech).

E. Results and Discussion

Fig. 4 (a)-(b) demonstrates that the proposed SEA consistently shows improved PESQ scores over the benchmark SEAs, except the AKF-Ideal method for all noise conditions as well as the SNR levels. The IAM+IFD method [19] relatively exhibits better PESQ score among the benchmark methods across the noise experiments. The Noisy speech shows the worse PESQ score for all noise conditions.

Fig. 4 (c)-(d) also shows that the proposed method demonstrates a consistent QSTI score improvement across the noise experiments as well as the SNR levels, apart from the AKF-Ideal method. The existing IAM+IFD method [19] is found to be competitive with the proposed method in QSTI improvement typically at low SNR levels. However, at high SNR levels, all the SEAs, even the noisy speech signal relatively shows the competitive QSTI score for all noise conditions.

It can be seen that the proposed SEA (Fig. 5 (f)) exhibits significantly less residual noise in the enhanced speech than that of the benchmark SEAs (Fig. 5 (c)-(e)) and is closely similar to the AKF-Ideal method (Fig. 5 (g)). When going from RWF-FCN method [18] to IAM+IFD method [19] (Fig. 5 (c)-(e)), noise-flooring is seen decreasing. The informal listening tests conducted on the enhanced speech also confirm that the benchmark SEAs relatively produce annoying sound as compared to negligible audio artifacts by the proposed method.

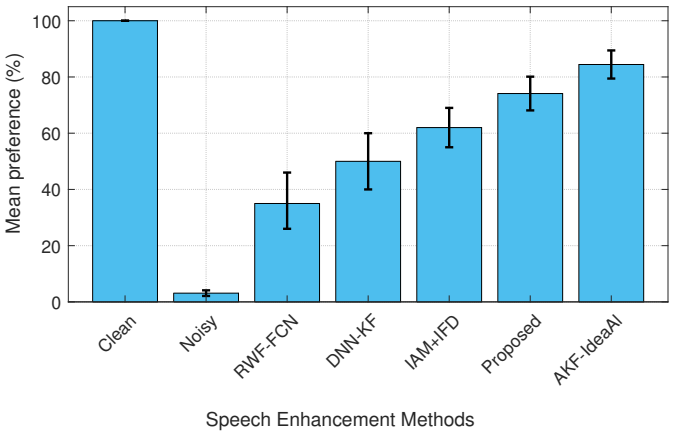

Fig. 6. The mean preference score (%) for each SEA on sp05 corrupted with 5 dB *passing car* noise.

The outcome of AB listening tests in terms of mean preference score (%) is shown in Fig. 6. It can be seen that the enhanced speech produced by the proposed SEA is widely preferred by the listeners (around 74%) than the benchmark methods, apart from the AKF-Ideal method (around 84%) and clean speech signal (100%). The IAM+IFD method [19] is found to be the best preferred (62%) amongst the benchmark methods, followed by the DNN-KF method [20] (50%), and RWF-FCN method [18] (35%).

V. Conclusion

This paper introduced a causal convolution encoder decoder-based augmented Kalman filter for speech enhancement in various noise conditions. Specifically, the proposed CCED network gives an estimate of the instantaneous noise magnitude spectrum to compute the noise PSD. Then the noise LPCs are computed from the estimated noise PSD. A whitening filter is also constructed with the estimated noise LPCs. It is employed to the noise corrupted speech, yielding a pre-whitened speech. The speech LPCs are computed from the pre-whitened speech. The large training set of CCED network enables the speech and noise LPC estimates to be effective in various noise conditions. As a result, the AKF constructed with the improved LPCs of speech and noise signal minimizes the residual noise as well as distortion in the enhanced speech. Extensive objective and subjective testing imply that the proposed method outperforms the benchmark methods in various noise conditions for a wide range of SNR levels.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211, April 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[5] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 629–632, May 1996.

[6] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 289–292, May 2004.

[7] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 177–180, April 1987.

[8] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, pp. 574 – 584, 2015.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[10] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, August 1991.

[11] S. K. Roy, W. P. Zhu, and B. Champagne, "Single channel speech enhancement using subband iterative kalman filter," *IEEE International Symposium on Circuits and Systems*, pp. 762–765, May 2016.

[12] A. E. W. George, S. So, R. Ghosh, and K. K. Paliwal, "Robustness metric-based tuning of the augmented Kalman filter for the enhancement of speech corrupted with coloured noise," *Speech Communication*, vol. 105, pp. 62 – 76, December 2018.

[13] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[14] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[15] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[16] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *Proceedings of Interspeech*, p. 1993–1997, 2017.

[17] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," *Proceedings of Interspeech*, pp. 3229–3233, 2018.

[18] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[19] N. Zheng and X. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 63–76, 2019.

[20] H. Yu, Z. Ouyang, W. Zhu, B. Champagne, and Y. Ji, "A deep neural network based Kalman filter for time domain speech enhancement," *IEEE International Symposium on Circuits and Systems*, pp. 1–5, May 2019.

[21] S. V. Vaseghi, "Linear prediction models," in *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, 2009, ch. 8, pp. 227–262.

[22] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," 2016.

[23] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[24] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," 2017.

[25] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[27] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," 2015.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, April 2015.

[29] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[31] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proceedings Interspeech 2010*, 2010, pp. 3110–3113.

[32] G. Hu, "100 nonspeech environmental sounds," *The Ohio State University, Department of Computer Science and Engineering*, 2004.

[33] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, "Smartphone-based real-time classification of noise signals using subband features and random forest classifier," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2204–2208.

[34] F. Saki and N. Kehtarnavaz, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 736–739.

[35] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[37] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, May 2001.

[38] B. Schwerin and K. K. Paliwal, "An improved speech transmission index for intelligibility prediction," *Speech Communication*, vol. 65, pp. 9–19, December 2014.

[39] K. K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.